

Geodesic-informed Generative Diffusion Model For Topology-preserved Image Video Generation

Nian Wu ¹, Nivetha Jayakumar ¹, Jiarui Xing ³, Miaomiao Zhang ^{1,2}

¹ University of Virginia, Department of Electrical & Computer Engineering, Charlottesville, VA, USA

² University of Virginia, Department of Computer Science, Charlottesville, VA, USA

³ Yale School of Medicine, New Haven, CT, USA

Abstract

Generative diffusion models have emerged as a class of powerful techniques for various imaging applications, including but not limited to synthesis, reconstruction, and segmentation. Despite their success, current generative models pose two key limitations. First, they primarily rely on image intensity and texture information, with limited attention to underlying object geometry. As a result, they do not guarantee geometric or topological consistency during the generation process, which is a crucial requirement for high-stakes domains such as computational anatomy, biology, and robotics, where preserving object structure is critical. Second, existing models fail to explicitly learn or represent shape changes in the generative process. Such deformation dynamics remain occluded within network parameters; hence leaving the transformation process uninterpretable and physically uninformed. To address these challenges, we introduce IGG (Image Generation informed by Geodesic dynamics), a novel framework that integrates topology-preserving geodesic principles into the diffusion-based generative process. In contrast to conventional methods that operate in image intensity space, IGG learns and synthesizes diverse samples within geodesic deformation spaces, where geometric object changes are learned as smooth and invertible smooth mappings from a given template/source image. Specifically, IGG employs a two-stage architecture: (i) a geodesic-informed image registration (GIR) module that directly encodes geodesic paths of image deformations into a compact latent space, and (ii) a latent geodesic diffusion (LGD) model that captures the distribution of these deformation representations, conditioned on a template image and/or text prompts. We validate IGG on the datasets of plant growth and brain MRI scans. Experimental results demonstrate that IGG outperforms state-of-the-art image generation and editing models, producing realistic, high-quality images with preserved topology and fewer artifacts. Furthermore, when used for data augmentation in downstream segmentation tasks, IGG substantially improves segmentation accuracy, particularly in low-data regimes. Our code is publicly available at <https://github.com/nellie689/IGG>.

Keywords

Geodesics, Diffeomorphisms, Generative diffusion model, Neural Operator

Article informations

<https://doi.org/10.59275/j.melba.2026-eec1>

©2026 Nian Wu, Nivetha Jayakumar, Jiarui Xing, and Miaomiao Zhang. License: CC-BY 4.0

Volume 2026, Received: 2025-11, Published 2026-08

Corresponding author: bsw3ac@virginia.edu



1. Introduction

Recent advances in deep generative modeling have significantly impacted various image analysis tasks, including but not limited to image synthesis (Xing et al., 2024; Xie et al., 2025; Liu et al., 2021; Reinhold et al., 2021), reconstruction (Jayakumar et al., 2023; Tezcan et al., 2018), translation (Deb et al., 2025; Graf et al., 2023; Özbey

et al., 2023), and more (Yang et al., 2019; Hossain and Zhang, 2025). Early research works, such as the variational autoencoder (VAE), assume a Gaussian prior over the latent feature space and learn its approximate posterior through variational inference (Kingma and Welling, 2013). Although effective, these models can produce overly smooth or blurry outputs due to their simplified assumptions (Hossain and Zhang, 2025). Another category of generative

modeling is the generative adversarial networks (GANs), which learn to produce realistic data samples by training a generator to compete against a discriminator that distinguishes generated samples from real data (Goodfellow et al., 2020). However, such models may suffer from mode collapses, where the generator produces limited diversity of outputs by ignoring other modes of the true data distribution (Zhang et al., 2018; Bang and Shim, 2021; Li et al., 2021). In contrast, diffusion models have recently emerged as a powerful class of deep generative frameworks that achieve state-of-the-art performance in modeling complex, multimodal data distributions and generating photorealistic images across diverse applications (Mukhopadhyay et al., 2023; Ho et al., 2020; Deb et al., 2025). These models employ an implicit generative scheme by learning to reverse a gradual noising process and iteratively transforming random noise into fine-grained and high-fidelity image samples (Ho et al., 2020). Due to these capabilities, diffusion models have demonstrated remarkable success in synthesizing highly realistic medical images to address critical challenges in biomedical research, such as data/label scarcity (Sharma et al. (2023); Jayakumar et al. (2024)), missing or incomplete modalities (Jayakumar et al. (2023); Deb et al. (2025)), and the need for predictive imaging to support disease progression modeling and clinical decision-making (Yoon et al. (2023); Dao et al. (2024); Zhang et al. (2024)).

Despite the success of the aforementioned generative diffusion models, current approaches face several limitations that can constrain their practical use in high-stakes domains requiring high accuracy and fidelity, such as medical imaging, robotics, and autonomous systems. In particular, most existing diffusion models focus on image intensity and texture, with little explicit consideration of object geometry during learning and generation (Nguyen et al., 2024; Li et al., 2024; Hertz et al., 2022; Kavar et al., 2023). As a result, they may produce unrealistic or biologically implausible samples, which poses significant risks for downstream applications where accurate geometric preservation is critical (Azizi et al., 2023; Jayakumar et al., 2023; Starck et al., 2025). Recent advances have introduced simple geometric constraints such as shape localization (Patashnik et al., 2023; Gupta et al., 2024), boundary conditions (Mazé and Ahmed, 2023), and 3D shape priors (Yu et al., 2025; Hu et al., 2024) in the modeling process. However, these approaches lack fine-grained structural preservation that is highly desirable in sensitive domains, including computational anatomy and medical imaging. Another important yet under-investigated limitation of current diffusion models is their inability to provide interpretable metrics or quantitative measures of topological integrity for image objects. Commonly used evaluation metrics include Fréchet Inception Distance (FID) for assessing distribution similarity between real and generated data (Heusel et al.,

2017), Inception Score (IS) for measuring diversity and quality (Salimans et al., 2016), and Structural Similarity Index (SSIM) for evaluating pixel-level fidelity (Wang et al., 2004). While these metrics effectively assess visual quality, they do not account for the preservation of object geometry and topology, raising questions about their validity in structure-sensitive applications.

In addition, existing diffusion models do not explicitly capture or represent object shape changes during the generative process (Ho et al., 2020; Dao et al., 2024; Jayakumar et al., 2024; Deb et al., 2025; Yoon et al., 2023). Recent studies have demonstrated that incorporating a geodesic-constrained deformation process to model object shape changes within the network can promote physically interpretable deformation trajectories, which enables the registration framework to capture the underlying dynamics of shape evolution rather than performing purely data-driven alignment (Wu and Zhang, 2023, 2024). Moreover, this geodesic guidance provides an additional source of regularization in the deformation learning process; thereby improving the biological plausibility of the registration results by suppressing artifacts such as folding, tearing, and crossing in the predicted transformations.

This paper presents an extended work of *Image Generation informed by Geodesic dynamics (IGG)* (Wu et al., 2025), that for the first time generates images by deforming a given template along random geodesics in deformation spaces guided by text instructions. In contrast to current approaches (Brooks et al., 2023; Ho et al., 2022; Kim et al., 2022), the IGG framework ensures topological consistency by treating each generated image as a deformed variation of a template or reference image through learned diffeomorphic transformations (also known as a one-to-one, smooth, and invertible mapping). A dynamic geodesic, producing natural transformations and providing interpretable metrics to quantify topological changes, are explicitly learned during the diffusion process. More specifically, the IGG model comprises two key components: a latent representation learning framework of diffeomorphic transformations informed by geodesic dynamics, and a latent geometric diffusion model conditioned on user-defined text instructions. Building on top of IGG (Wu et al., 2025), our work further

- (i) Expand the core method section of IGG by introducing details of the geodesic-informed image registration module, which provides a more principled representation of topology-preserving transformations.
- (ii) Thoroughly investigate the quantitative metrics to assess both the fidelity and topological consistency of the generated samples.
- (iii) Design downstream experiments in the context of image segmentation to demonstrate an additional clinical utility

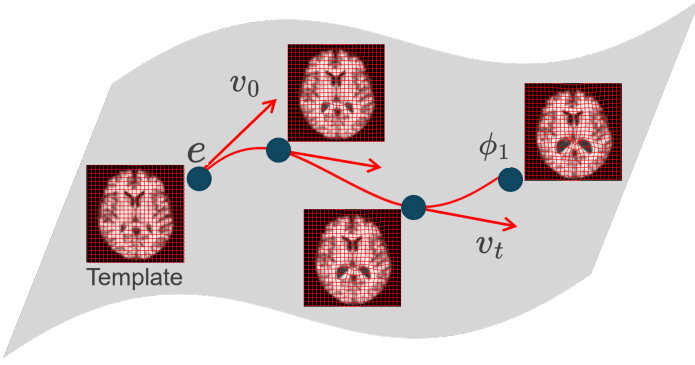


Figure 1: An illustrated example of a geodesic path in the space of brain deformations.

of the generated samples in real-world applications.

We demonstrate the effectiveness of IGG on a diverse set of real-world datasets, including Komatsuna plant growth data (Uchiyama et al., 2017) and brain MRIs (LaMontagne et al., 2019). Experimental results show that IGG significantly outperforms state-of-the-art text-guided generative models (Ho et al., 2022; Yang et al., 2024; Xing et al., 2025) in producing image samples with well-preserved topological structures. We further demonstrate the potential of IGG to enhance deep learning-based downstream tasks by training brain segmentation networks with synthesized image-label pairs. It is worth noting that, once trained, IGG generalizes well across image domains (e.g., variations in image contrast and scanner types) due to its flexibility in generating geometric deformations conditioned on any given template.

2. Background: Geodesics In Deformation Spaces

In this section, we briefly review the basic concept of geodesics in deformation spaces, which provides a principled way to model smooth and natural deformations between images while preserving structural integrity (Miller et al., 2002; Beg et al., 2005). With the underlying assumption that objects of a generic class (e.g., human brains, hearts, or lungs) are described as deformed variants of a given template, descriptors of that class arise naturally by deforming the template to other images along a **geodesic** - a shortest path with locally minimized energy of transformations (Avants et al., 2008; Joshi et al., 2004). In theory, every topological property of the deformed template can be preserved by enforcing the resulting transformations to be diffeomorphisms, i.e., differentiable, bijective mappings with differentiable inverses (Beg et al., 2005; Arnold, 1966; Miller et al., 2002). Violations of such constraints introduce image artifacts, such as tearing, crossing, or passing through itself.

Let $\text{Diff}^\infty(\Omega)$ denote the space of smooth diffeomorphisms on a d -dimensional torus domain $\Omega = \mathbb{R}^d/\mathbb{Z}^d$. The tangent space of diffeomorphisms is the space $V = \mathcal{X}^\infty(T\Omega)$ of smooth vector fields on Ω . Consider a time-varying velocity field, $\{v_t\} : [0, 1] \rightarrow V$, we can generate diffeomorphisms $\{\phi_t\}$ between pairwise images by solving

$$\frac{d\phi_t}{dt} = v_t(\phi_t), t \in [0, 1], \quad (1)$$

where ϕ_0 is the identity map, e , and ϕ_1 is the target transformation.

The geodesic minimizes the functional $\int_0^1 (\mathcal{L}v_t, v_t) dt$ (Beg et al., 2005), where $\mathcal{L} : V \rightarrow V^*$ is a symmetric, positive-definite differential operator that maps a tangent vector $v(t) \in V$ into its dual space as a momentum vector $m(t) \in V^*$. We typically write $m(t) = \mathcal{L}v(t)$, or $v(t) = \mathcal{K}m(t)$, with $\mathcal{K}$ being an inverse operator of $\mathcal{L}$. In this paper, we adopt a commonly used Laplacian operator $\mathcal{L} = (-\alpha\Delta + e)^3$, where α is a weighting parameter that controls the smoothness of transformation fields. The $(\cdot, \cdot)$ is a dual pairing, which is similar to an inner product between vectors. An example of geodesic path in the space of deformations is illustrated in Fig. 1.

According to a well-known geodesic shooting algorithm (Vialard et al., 2012), the minimum of the functional mentioned above is uniquely determined by solving a Euler-Poincaré differential (EPDiff) equation (Vialard et al., 2012; Younes et al., 2009) with a given initial condition. That is, for $\forall v_0 \in V$, a geodesic path $t \mapsto \phi_t$ in the space of diffeomorphisms can be computed by forward shooting the EPDiff equation

$$\frac{\partial v_t}{\partial t} = -K \left[(Dv_t)^T m_t + Dm_t v_t + m_t \text{div } v_t \right], \quad (2)$$

where D denotes the Jacobian matrix and div is a divergence operator.

Derive geodesics from images. Consider a source image S and a target image T defined in the domain Ω ($S(x), T(x) : \Omega \rightarrow \mathbb{R}$). The optimization of geodesic transformations can be formulated as minimizing an energy function over the initial velocity, subject to the EPDiff equations, i.e.,

$$E(v_0) = (\mathcal{L}v_0, v_0) + \lambda \text{Dist}(S(\phi_1), T) \text{ s.t. Eq. (1) \& (2)}.$$

Here, $\text{Dist}(\cdot, \cdot)$ is a distance function that measures the dissimilarity between images, and λ is a positive weighting parameter. In this paper, we will use the commonly used sum-of-squared intensity differences (Beg et al., 2005; Wu et al., 2024; Zhang et al., 2017).

3. Method: IGG

This section introduces an extended work of, IGG, that for the first time generates images in deformation spaces guided by text instructions (Wu et al., 2025). A dynamic geodesic, producing natural transformations and providing interpretable metrics to quantify topological changes, will be explicitly learned during the generative diffusion process. Our IGG model consists of two main components: (i) a geodesic-informed image registration (GIR) network that directly encodes geodesic paths of image deformations into a compact latent space; and (ii) a latent geodesic diffusion (LGD) model that captures the latent distribution of these deformation representations, conditioned on a template image and/or user-defined text prompts. An overview of our model is illustrated in Fig. 2.

3.1 Representation Learning of Geodesic Deformations via GIR

Consider a given set of N template/source and target image pairs associated with text instructions $\{s^n, f^n, \Lambda^n\}_{n=1}^N$. We introduce a GIR network, based on an autoencoder architecture, to learn the latent representation of geodesic transformations, parameterized by velocity fields $\{v_t\}$, between a template s^n and a target f^n . The encoder of GIR, $\mathcal{E}_{\theta_v}$, maps input data to the latent representation of a sequence of velocity fields, $\{z_{v_t}\}_{t=0}^1$. Inspired by the recent work NeurEPDiff (Wu and Zhang, 2023), which develops a neural operator to learn mappings governed by the EPDiff equation, we incorporate the NeurEPDiff module, $\mathcal{G}_{\theta_r}$, into our IGG model to learn geodesics in latent deformation spaces. More specifically, starting with a latent initial velocity z_{v_0} , the operator $\mathcal{G}_{\theta_r}$ will iteratively propagate it along the geodesic path, $z_{v_0} \mapsto \dots z_{v_t} \mapsto \dots z_{v_1}$, through a designed latent geodesic evolution module.

Latent Geodesic Evolution. Similar to (Wu and Zhang, 2023, 2024), we leverage neural operators as surrogate models to approximate the mapping function of the EPDiff equation during network training. In particular, we employ a multilayer neural network with J hidden layers to simulate the geodesic mapping function (EPDiff) from z_{v_t} to $z_{v_{t+1}}$. For notational simplicity, we let u^j denote the input of the j -th hidden layer and u^{j+1} its output, where $j \in \{1, \dots, J\}$. At each hidden layer, we combine a *local linear transformation* W^j with a *global convolutional kernel* $\mathcal{H}^j$, enabling the extraction of both *local* and *global representations*. A nonlinear activation function, $\sigma(\cdot)$, is defined as a Gaussian Error Linear Unit (GeLU) (Hendrycks and Gimpel, 2016) with a smoothing operator K to ensure continuity and smoothness in the output signal, i.e., $\sigma(\cdot) := K(\text{GeLU}(\cdot))$. Formally, we update $u^{j+1} \leftarrow u^j$ by

$$u^{j+1} := \sigma(W^j u^j + \mathcal{H}^j * u^j),$$

where $*$ represents a convolution operator. Note that, for computational efficiency, the global convolution is implemented as an element-wise multiplication in the Fourier domain, followed by an inverse transform back to the spatial domain.

Finally, the geodesic evolution of the latent velocity from z_{v_t} to $z_{v_{t+1}}$ at each time point t can be defined as

$$z_{v_{t+1}} := \sigma(W^J, \mathcal{H}^J) \circ \dots \circ \sigma(W^2, \mathcal{H}^2) \circ \sigma(W^1, \mathcal{H}^1, z_{v_t}),$$

where $\circ$ denotes the composition of network operations. The decoder, $\mathcal{D}_{\theta_v}$, then projects these latent representations back to the full-dimensional image space to produce a geodesic flow of velocity fields, $v_0 \mapsto \dots v_t \mapsto \dots v_1$. The corresponding deformations, $\phi_0 \mapsto \dots \phi_t \mapsto \dots \phi_1$ are then generated according to Eq. 1. For simplicity of notation, we define $\Phi \triangleq \{\phi_t\}_{t=0}^1$ in the following sections.

The network loss combines contributions from an unsupervised registration loss, and a geodesic loss guided by numerical solutions of the EPDiff equation obtained via Euler integration. These solutions are represented as $\{\hat{v}_t^n\}, t \in (0, 1]$ with a simultaneously learned initial velocity $\hat{v}_0^n$. To simplify the notation, we define $\hat{v}^n \triangleq \{\hat{v}_t^n\}$ before formulating the loss function as

$$\begin{aligned} \mathcal{L}(\theta_v, \theta_r) = & \sum_{n=1}^N \lambda \|(s^n(\phi_1^n(\theta_v)) - f^n\|_2^2 \\ & + \frac{1}{2}(\mathcal{L}v_0^n(\theta_v), v_0^n(\theta_v)) \\ & + \eta \|\mathcal{D}_{\theta_v}(\mathcal{G}_{\theta_r}(z_{v_0})) - \hat{v}^n\|_2^2 \\ & + \text{Reg}(\theta_r, \theta_v), \quad s.t. \text{ Eq. (1)}, \end{aligned} \quad (3)$$

where λ and η are positive weighting parameters to balance the image matching term and the geodesic loss, and $\text{Reg}(\cdot)$ is a network regularization.

3.2 Latent Geodesic Diffusion Model

With a pre-trained GIR network, we are now ready to introduce a latent geodesic diffusion module to simulate the distribution of latent geodesic flows for image transformations, $\{z_{v_t}^n\}_{t=0}^1$. To simplify the notation, we define $z^n := \{z_{v_t}^n\}_{t=0}^1$ for the n -th subject, where all time-dependent components $\{z_{v_t}^n\}$ are concatenated to form a unified representation. In contrast to previous approaches that perform the diffusion process in the image space (Brooks et al., 2023; Ho et al., 2022; Kim et al., 2022), our latent geometric diffusion operates in the deformation space. Specifically, our diffusion process directly samples geodesics of topology-preserving diffeomorphic transformations, which are then applied to deform the input template to generate the final output.

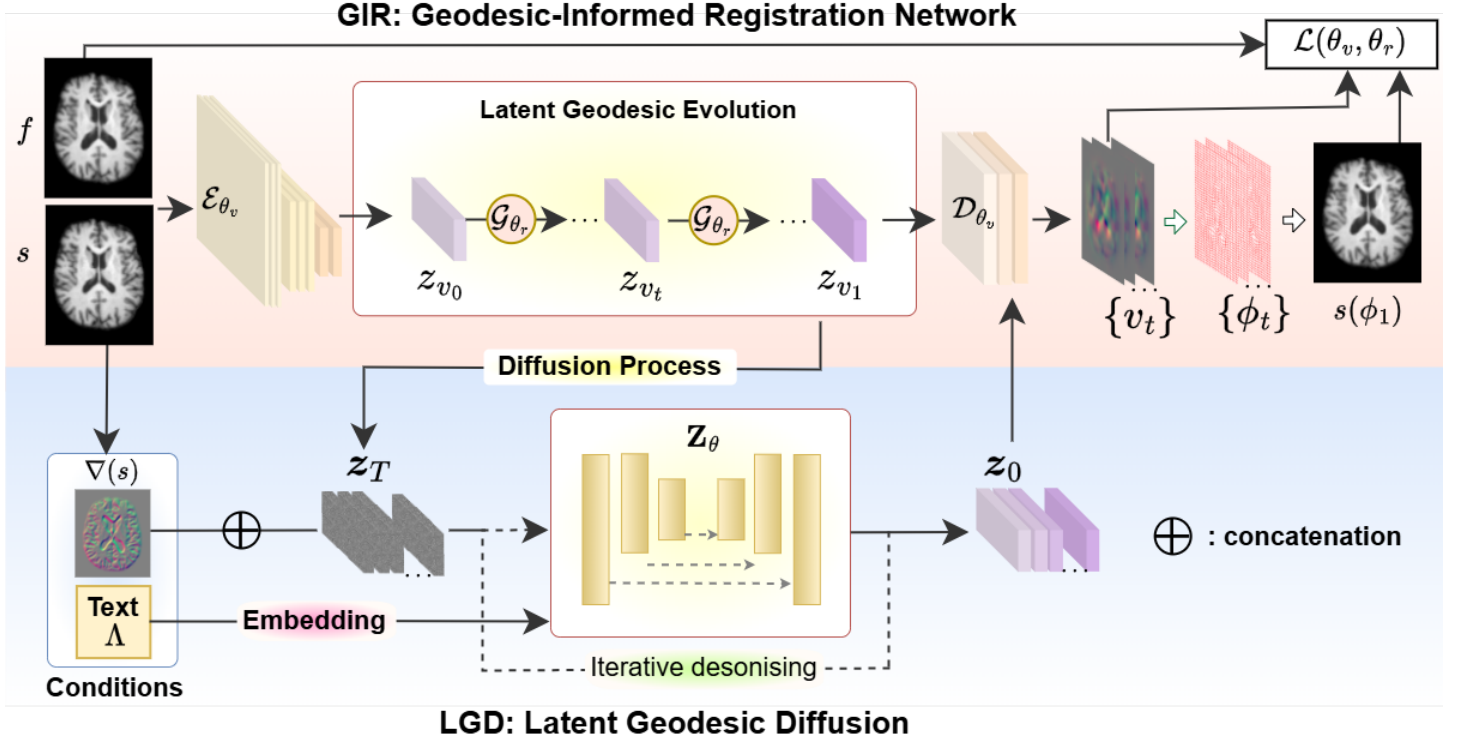


Figure 2: An overview of our IGG framework with two key components: (i) GIR to learn the latent representations of geodesic transformations from paired template and target images; and (ii) LGD to generate a time-sequence of image transformations over time, conditioned on text instructions.

3.2.1 Forward diffusion in geodesic deformation spaces.

Following the principles of denoising diffusion probabilistic models (DDPM) (Ho et al., 2020) and video diffusion models (Ho et al., 2022), our forward diffusion is fixed to a Markov chain that progressively adds noise, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, to the initial latent geodesic representations $z_0^n \sim p(z_0)$, with a number of T steps. Mathematically, the forward diffusion process can be formulated as

$$q(z_\tau^n | z_{\tau-1}^n) := \mathcal{N}(z_\tau^n; \sqrt{1 - \beta_\tau} z_{\tau-1}^n, \beta_\tau \mathbf{I}),$$

where $\tau \in [1, \dots, T]$ and β_τ is a time-dependent variance schedule of scalar values $\in (0, 0.999]$ used to parameterize the probabilistic transitions. Alternatively, the forward diffusion process of sampling z_τ at an arbitrary timestep τ can be formulated in closed form using

$$\begin{aligned} q(z_\tau^n | z_0^n) &= \mathcal{N}(z_\tau^n; \sqrt{\bar{\alpha}_\tau} z_0^n, (1 - \bar{\alpha}_\tau) \mathbf{I}), \\ z_\tau^n &= \sqrt{\bar{\alpha}_\tau} z_0^n + \sqrt{1 - \bar{\alpha}_\tau} \epsilon, \end{aligned} \quad (4)$$

where $\alpha_\tau = 1 - \beta_\tau$ and $\bar{\alpha}_\tau = \prod_{s=1}^{\tau} \alpha_s$, where $s \in [1, \dots, \tau]$.

3.2.2 Reverse diffusion with geometric conditioning.

Similar to Ho et al. (2020), the reverse diffusion process iteratively denoises the latent variable of the diffusion model z_τ^n

over τ steps, given template-specific conditions. We introduce a novel geometric conditioning in the latent geodesic space, which enables the model to adapt more effectively to specific contexts; hence improving its generative performance based on observed template images and provided text instructions. More specifically, our geometric conditioning is achieved by concatenating two components: image embedding in the form of downsampled template image gradient, $\{\nabla s^n\}$ and the geodesic deformation embedding, represented by the latent velocity fields, $\{z_\tau^n\}$. Following a similar principle as seen in (Ho et al., 2020), we sample the denoised latent geodesic flow of velocity fields z_0^n , starting from the noisy state z_τ^n at the τ -th time-step and progressively remove noise to reconstruct the original latent representation. This process is mathematically expressed as

$$\begin{aligned} p_\theta(z_{\tau-1}^n | z_\tau^n, \nabla s^n, \Lambda^n, \tau) &= \\ \mathcal{N}(z_{\tau-1}^n; \mu_\theta(z_\tau^n, \nabla s^n, \Lambda^n, \tau), \Sigma_\theta(z_\tau^n, \nabla s^n, \Lambda^n, \tau)), \end{aligned}$$

where μ_θ and Σ_θ represent the parameterized mean and variance of the approximated Gaussian distribution at each time step of the reverse diffusion process (Ho et al., 2022). In our experiments, we employ a 3D UNet architecture (Ronneberger et al., 2015) Z_θ as the deep predictive model, which predicts the noise to be removed

from z_τ^n to reconstruct $z_{\tau-1}^n$ using the equation:

$$z_{\tau-1}^n = \frac{1}{\sqrt{\alpha_\tau}}(z_\tau^n - \frac{1 - \alpha_\tau}{\sqrt{1 - \bar{\alpha}_\tau}} \mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \Lambda^n, \tau)) + \sigma_t I, \quad (5)$$

where $I \sim \mathcal{N}(0, \mathbf{I})$ and $\sigma_t = \tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \cdot \beta_t$.

The text embedding, encoded using a pre-trained CLIP text encoder (Radford et al., 2021), $\{\Lambda^n\}$ is integrated into this network via cross-attention. With an optimized denoising network $\mathbf{Z}_\theta$, the iterative sampling using Eq 5 yields a clean geodesic deformation embedding z_0^n .

The network loss of our proposed latent geodesic diffusion module is formulated as

$$\mathcal{L}_\theta = \frac{1}{N} \sum_{n=1}^N \|\epsilon_\tau^n - \mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \Lambda^n, \tau)\|_2^2 + \text{reg}(\theta), \tau \in \mathcal{U}[1, T], \quad (6)$$

where $\text{reg}(\cdot)$ is a regularization of the network parameter θ , $\mathcal{U}$ denotes a uniform distribution and T is the total number of diffusion timesteps.

Classifier-free guidance. To balance the trade-off between sample quality and diversity of velocity fields in the latent space after training, we optimize the combination of conditional and unconditional diffusion models. This is achieved by leveraging image-conditioned guidance with scale δ_I , text-conditioned guidance with scale δ_T , and a null-condition $\emptyset$. Following similar principles from the classifier-free guidance approach (Ho and Salimans, 2022), we first jointly train a conditional and an unconditional IGG model. The resulting score estimates from these models are then combined to predict the noise $\hat{\mathbf{Z}}$ as

$$\begin{aligned} & \hat{\mathbf{Z}}_\theta(z_\tau^n, \nabla s^n, \Lambda^n, \tau) \\ &= \mathbf{Z}_\theta(z_\tau^n, \emptyset, \emptyset, \tau) \\ &+ \delta_I \cdot [\mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \emptyset, \tau) - \mathbf{Z}_\theta(z_\tau^n, \emptyset, \emptyset, \tau)] \\ &+ \delta_T \cdot [\mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \Lambda^n, \tau) - \mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \emptyset, \tau)] \\ &= (1 - \delta_I) \cdot \mathbf{Z}_\theta(z_\tau^n, \emptyset, \emptyset, \tau) \\ &+ (\delta_I - \delta_T) \cdot \mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \emptyset, \tau) \\ &+ \delta_T \cdot \mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \Lambda^n, \tau). \end{aligned}$$

3.3 Network Optimization

We first optimize the loss of representation learning of geodesic transformations (Eq. (3)), followed by the latent geodesic diffusion model (Eq. (6)). The training process of IGG is summarized in Alg. 1. During testing, the procedure for sampling the geodesic of latent velocity fields, conditioned on a given template image and text instructions, is outlined in Alg. 2.

Algorithm 1 IGG Training

Input: Data $\{s^n, f^n, \Lambda^n\}_{n=1}^{N_{\text{train}}}$

- 1: Pretrain $\mathcal{E}_{\theta_v}, \mathcal{E}_{\theta_r}, \mathcal{D}_{\theta_v}$ via Eq. (3)
 - 2: **repeat**
 - 3: Initial latent geodesic representations, $z_0^n \leftarrow \mathcal{G}_{\theta_r}(\mathcal{E}_{\theta_v}(s^n, f^n))$,
 - 4: Sample $\tau \sim \mathcal{U}(1, T)$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$,
 - 5: Sample z_τ^n at τ by $z_\tau^n \leftarrow q(z_\tau^n | z_0^n)$ via Eq. (4),
 - 6: Optimize $\mathbf{Z}_\theta$ by taking a gradient descent step on Eq. (6).
 - 7: **until** converged
-

Algorithm 2 IGG Sampling

Input: Data $\{s^n, \Lambda^n\}_{n=1}^{N_{\text{test}}}$

Output: Geodesic deformations Φ^n

- 1: Initialize $z_T^n \sim \mathcal{N}(0, \mathbf{I})$
 - 2: **for** $\tau = T, \dots, 1$ **do**
 - 3: $\hat{z}_{\tau-1}^n \leftarrow p_\theta(z_{\tau-1}^n | z_\tau^n, \nabla s^n, \Lambda^n, \tau)$ via Eq. (5),
 - 4: **end for**
 - 5: $\{\hat{v}^n\} \leftarrow \mathcal{D}_{\theta_v}(\hat{z}_0^n)$,
 - 6: $\Phi^n \leftarrow \{\hat{v}^n\}$ via Eq. (1),
 - 7: **return** Φ^n
-

4. Experimental Evaluation

We evaluate the proposed IGG framework using a diverse set of real-world image datasets that capture deformable shape changes over time. We first assess the quality of the learned latent representations of geodesics by comparing them to numerical solutions of the EPDiff equation obtained via Euler integration.

Next, we evaluate the quality of the generated images deformed by the sampled geodesics guided by given text instructions. These results are compared with state-of-the-art generative models that synthesize image sequences/videos with publicly available training code or fine-tuning options, including the video diffusion model (VDM) (Ho et al., 2022), CogVideoX (Yang et al., 2024), and DynamiCrafter (Xing et al., 2025). Note that all baselines are trained using sequences of deformed images generated along the geodesic transformations derived from the numerical solutions of EPDiff equation. This approach ensures that the results are independent of the learned geodesics predicted by the first module of the IGG framework, providing a consistent and unbiased evaluation.

Finally, to further assess the quality of the generated samples and demonstrate their clinical utility in real-world applications, we design a downstream task in the context of brain MRI segmentation. More specifically, we use the synthesized image samples as augmentation data by pairing each transformed image with its corresponding segmenta-

tion label, both generated through the learned deformation fields applied to existing images and labels. We conduct experiments using a diverse set of widely used segmentation architectures, including U-Net (Ronneberger et al., 2015), TransUNet (Chen et al., 2021), and Attention U-Net (At-tUNet) (Oktay et al., 2018), under limited training data conditions.

4.1 Dataset

Komatsuna plant. We include 300 frames of RGB-D label-maps representing five Komatsuna plants from the publicly available data repository (Uchiyama et al., 2017). The label maps cover different leaves that emerge from the bud and grow in size over time, where the plant growth was monitored between 228-236 hours. All data frames were resampled to 128^2 and pre-aligned with affine transformations.

Longitudinal brain MRI. We include a total of 2618 T1-weighted longitudinal (time-series) brain MRIs sourced from the Open Access Series of Imaging Studies (OASIS-3) dataset (LaMontagne et al., 2019). This experiment aims to validate our method using longitudinal data that includes scans at varying time intervals for individuals spanning both healthy subjects and Alzheimer’s diseases (AD), aged 60-90. Given the scenario that many existing image generation/editing methods with text instructions focus on 2D natural images (Brooks et al., 2023; Meng et al., 2021), we specifically utilize 2D scans derived from this 3D brain data for comparison with state-of-the-art baselines. All MRIs undergo pre-processing, including resizing to 128^2 , with isotropic voxels of 1mm^2 , skull-stripping, intensity normalization, bias field correction, and pre-alignment using affine transformations.

Text instructions. Our text condition consists of a description of the template image, followed by details of the specific edits or progressions applied to it. For brain MRIs, this includes biological variables such as age, sex, and gender. In contrast, for plants, it primarily focuses on the growth timeline. All information is sourced directly from the original dataset available in the data repository.

4.2 Experimental Design and Implementation Details

Evaluate learned latent representations of geodesics via GIR. We evaluate GIR’s ability to predict geodesic dynamics by comparing the learned geodesics with numerical solutions to the EPDiff equation using Euler integration. The assessment focuses on errors in velocity fields, transformations, and deformed images between the two approaches. Quantitative metrics, such as the mean absolute error of velocity fields, are computed at each time integration step.

Evaluate generated samples. We compare the quality of samples generated by IGG with three baseline models: VDM (Ho et al., 2022), CogVideoX (Yang et al., 2024), and DynamiCrafter (Xing et al., 2025)). In particular, we first assess individual synthesized images at each time step using standard image synthesis metrics, including the Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) to measure the distributional similarity between generated and real images. Additionally, we use the Structural Similarity Index (SSIM) (Wang et al., 2004) to evaluate structural, contrast, and luminance similarities. We then evaluate the continuity and smoothness of generated video sequences by computing the Fréchet Video Distance (FVD) (Unterthiner et al., 2018) and providing qualitative visual comparisons. Finally, we assess the perceptual quality of the generated samples using the Inception Score (IS) (Salimans et al., 2016).

Evaluate the preservation of object geometry and topology. To demonstrate the efficiency of our IGG model in preserving fine-grained geometric structure and topology, we first generate image sequences by deforming a template image using the sampled geodesic flow of deformations produced by IGG and compare these results with all baseline models. Secondly, we present the corresponding determinant of the Jacobian (DetJac) maps for the deformations. The DetJac values reveal important patterns of volume change: a value of 1 indicates no volume change, $\text{DetJac} < 1$ reflects volume shrinkage, and $\text{DetJac} > 1$ implies volume expansion. A DetJac value below zero indicates artifacts or singularities in the transformation field, highlighting a failure to maintain the topological integrity. Note that only our IGG model enables such a metric to quantify the topological changes in the generated samples.

Evaluate reliability of model predictions. We evaluate the reliability and confidence of IGG in learning the geodesic deformations of geometric shapes over time by computing pixel-wise mean and standard deviation for individual sampled image sequences along the geodesic path. We visualize confidence intervals to highlight regions with 95% certainty in growth patterns. These intervals are defined by the mean of 1000 generated samples with pixel-wise bounds that are two standard deviations above and below the mean.

Demonstrate utility of generated samples in downstream segmentation tasks. We evaluate the effectiveness of IGG generated images by using them as augmented data for segmentation tasks under conditions of limited number of training samples. We then compare the performance of segmentation models trained with and without our

synthesized augmented data, considering dataset augmentations of $2\times$ and $4\times$ the size of the original dataset. The segmentation accuracy is assessed using the dice score (Dice, 1945), both averaged across all anatomical regions and separately for different structures. To highlight the flexibility of IGG in generating new data, we evaluate augmentation performance on OASIS-3 (LaMontagne et al., 2019) and OASIS-1 (Marcus et al., 2007) datasets independently, noting that our model is trained exclusively on OASIS-3. This setup allows us to examine the generalization capability of the synthesized data to an unseen dataset while quantifying its impact on segmentation performance.

Parameter Setting. We split all datasets into 80%, 10%, and 10% for training, validation, and testing. All experiments are conducted on NVIDIA A100 GPUs. For training the autoencoder for geodesic learning, we alternatively update the encoder-decoder and the latent geodesic neural operator, and finally jointly train both components for 2000 epochs per stage. We set the weight decay as $1e^{-4}$, batch size as 64, and a learning rate as $5e^{-4}$. For training the

latent geodesic diffusion model, we set the learning rate as $1e^{-4}$, the number of epochs as 6000, batch size as 36, and diffusion steps as 500. All models are trained using the Adam optimizer (Kingma and Ba, 2014).

4.3 Experimental Results

The left panel of Fig. 3 visualizes the predicted transformations, velocity fields, and deformed images from IGG’s geodesic-informed registration network (GIR), compared with numerical solutions to the EPDiff equation (Eq. (2)). These methods show a high degree of similarity. The right panel of Fig. 3 presents quantitative results of the absolute per-pixel error, computed between the velocity field obtained from the numerical solution of the EPDiff equation and the velocity field predicted by the GIR module at each time step. These results collectively highlight GIR’s ability to effectively learn geodesic mapping functions comparable to real numerical solutions.

We also report the additional computational overhead introduced by the GIR, which primarily comes from the

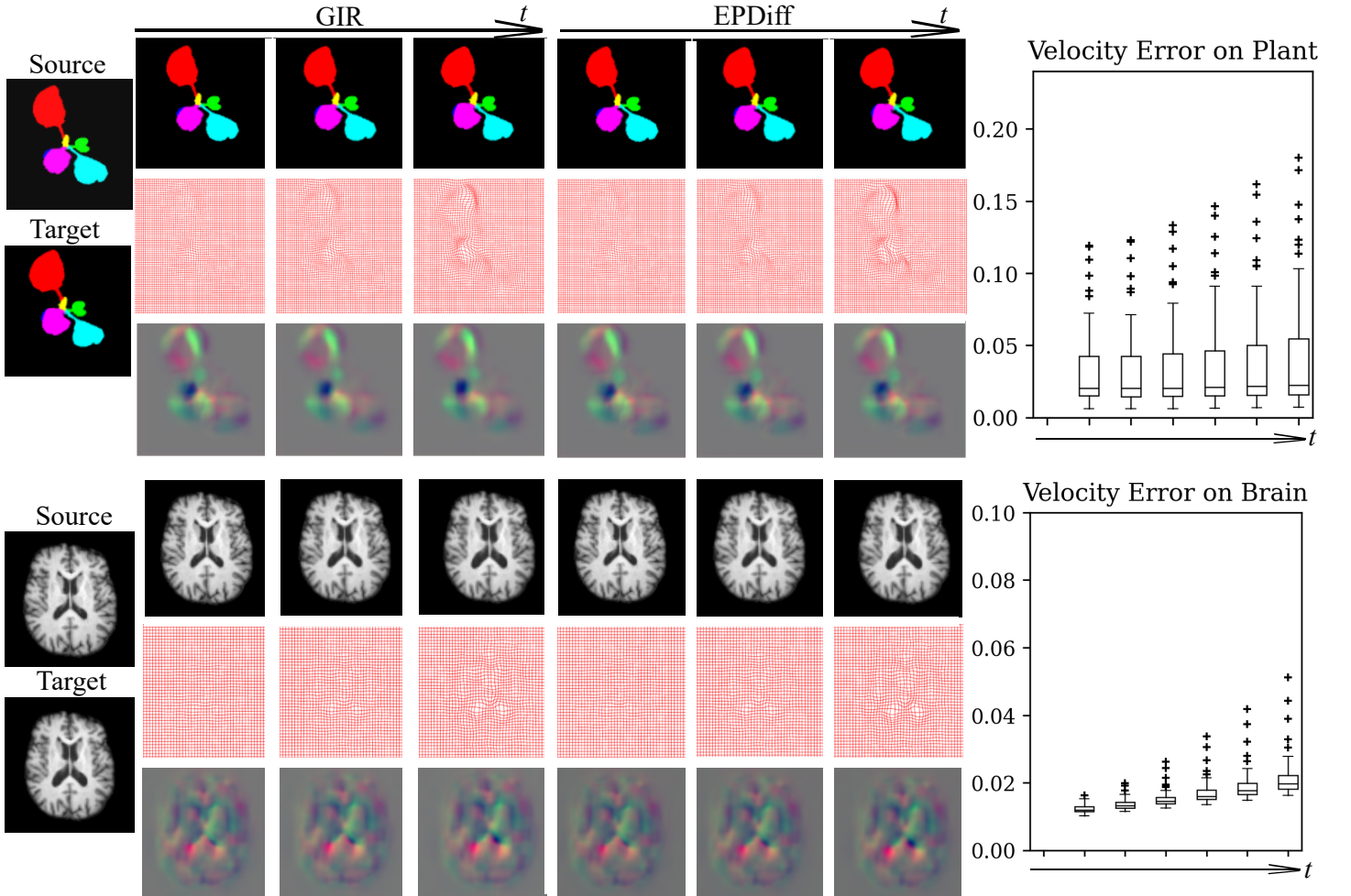


Figure 3: Comparison of predicted geodesics by GIR (the first sub-module of IGG) vs. real numerical solutions from the EPDiff equation. Left: Visualization of predicted deformed images, deformations, and velocities along time. Right: Mean absolute error of predicted velocities over time compared to numerical integration of EPDiff.

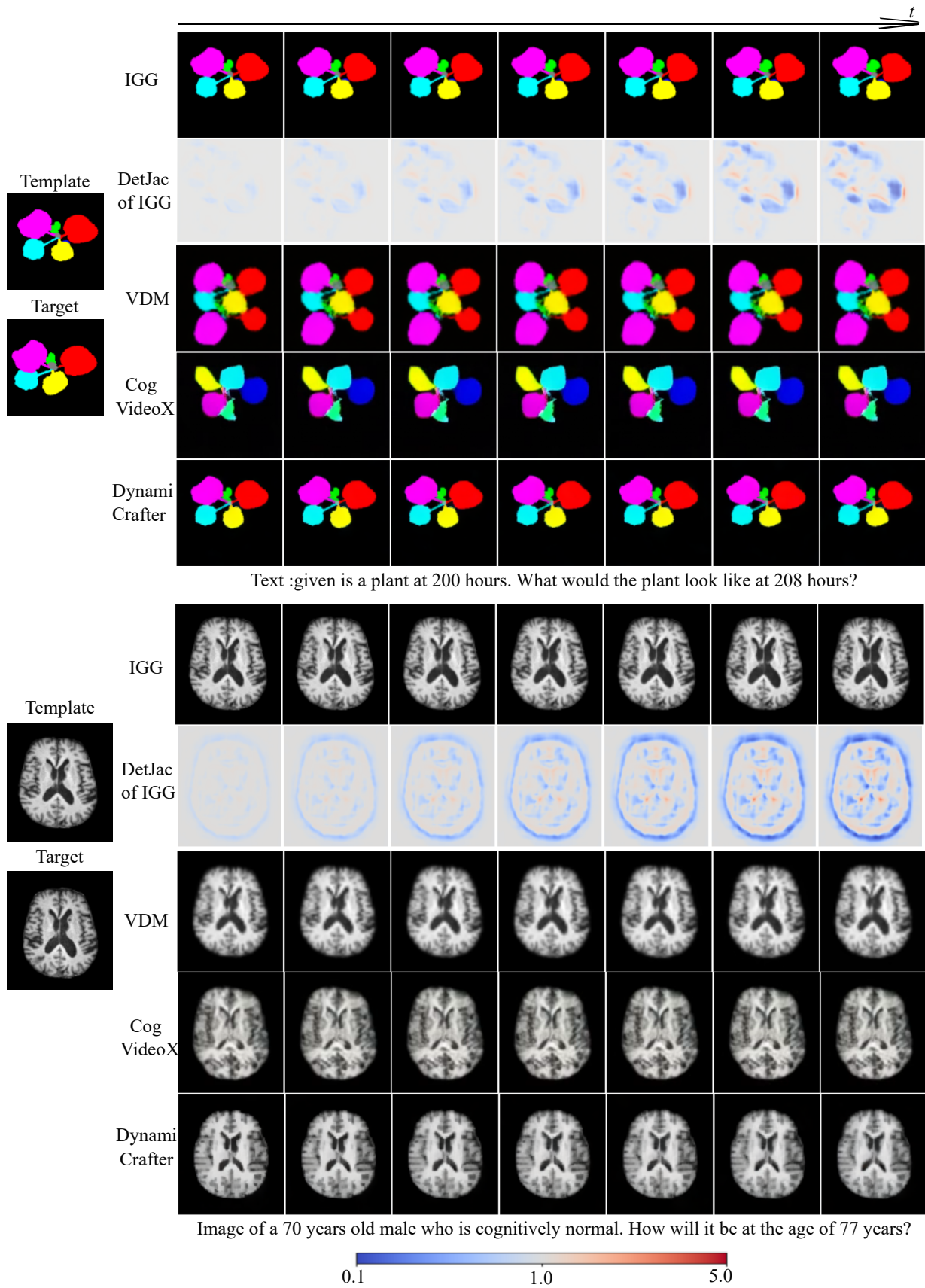


Figure 4: A comparison of images generated by IGG (with corresponding DetJac) against all baseline models across different time frames. Given an input template image and text instructions, all models generate samples of target images. The ground truth "target" along with input template images are provided on the left side of the panel for reference.

latent neural operator used to approximate the EPDiff equation (Eq. (2)). In practice, in our experiments on 128^2 brain images, the training time per epoch is approximately 13.10s for GIR, compared to 4.81s for a standard registration model with numerical solutions to the EPDiff Hinkle et al. (2018). The corresponding inference times are 0.01s vs. 0.004s, respectively.

Fig. 4 compares the generated images from our model IGG (along with the associated DetJac maps of the deformations) with the baseline generative models. The visualized images and DetJac values of IGG suggest that our model effectively preserves the topological structure of objects within the generated images with well-captured progression of geometric shape changes over time. In contrast, samples generated by the baselines fail to maintain the geometric integrity of various structures. For instance, in plant growth images, the leaves appear to merge or overlap unnaturally, disrupting their biological topology. Similarly, the baselines inaccurately predict parts of the ventricles, resulting in regions that do not correspond to the original brain structure. It is worth noting that the baseline methods perform diffusion directly in the image intensity space, learning the distribution of raw pixel values. During sampling, they generate pixel intensities through iterative denoising, which can lead to blurring (or noisy) artifacts observed in diffusion models operating in pixel space. In contrast, our method performs diffusion in the deformation space, generating velocity or deformation fields that are subsequently applied to a template image to synthesize new image sequences. By modeling structural variation through geometric transformations rather than pixel-wise generation, our approach better preserves sharp and anatomically coherent structures.

Tab. 1 reports the evaluation metrics of sample quality and diversity for the generated images from our model IGG and the baselines. While all models achieve similar IS, IGG shows significantly lower FVD, FID, and KID scores, and higher SSIM scores. This indicates that while all models generate diverse images (leading to a good IS), the baselines lack realistic features or proper distribution alignment

with real images. The observation of IGG achieving approximately 10 times better scores demonstrates its ability to generate more realistic images.

Fig. 5 visualizes examples of confidence maps of plant growth and brain progression images generated from IGG. It suggests that our model effectively captures the growth pattern of plants over time, primarily focusing on the boundaries of the leaves. The confidence maps of brain images highlight expanding patterns in ventricles, demonstrating the model’s ability to capture dynamic changes over time.

Fig. 6 presents the improvement in average dice scores for different segmentation models trained with varying dataset configurations. The baseline corresponds to training on the original limited dataset. It shows that with gradually increasing the amount of augmented data generated by IGG from $2\times$ to $4\times$ the size of the original training set, we can consistently obtain better segmentation performance with higher dice scores across all models. These results demonstrate that the generated image-label pairs are able to effectively augment real data, particularly under limited-label conditions. While the improvements on OASIS-1 (with maximal $+5.1\%$) are smaller than those on OASIS-3 (with maximal $+11.0\%$), they remain consistent, highlighting the generalizability of our model across datasets.

Fig. 7 further reports detailed dice scores across different anatomical structures when using IGG’s predictions as augmented data on both OASIS-3 and OASIS-1 datasets.

Fig. 8 visualizes representative segmentation maps predicted by a trained U-Net (Ronneberger et al., 2015) in different configurations of the training dataset. When trained on a limited dataset, and the model produces errors in certain anatomical structures, whereas training with augmented data yields predictions that more closely match the ground truth. These results further confirm the effectiveness of incorporating our generated samples, especially under limited-data conditions.

Table 1: A comparison of performance metrics across all methods for various datasets.

Dataset	Model	FVD ↓	FID ↓	KID ↓	SSIM ↑	IS ↑
Plant	IGG	48.29	9.23	0.007 ± 0.02	0.98 ± 0.02	1.06 ± 1.2%
	VDM (Ho et al., 2022)	357.56	71.79	0.23 ± 0.03	0.23 ± 0.10	1.05 ± 1.0%
	CogVideoX (Yang et al., 2024)	471.49	153.93	0.63 ± 0.07	0.69 ± 0.19	1.05 ± 0.6%
	DynamiCrafter (Xing et al., 2025)	466.70	80.78	0.41 ± 0.04	0.89 ± 0.04	1.03 ± 0.3%
	IGG	26.39	6.23	0.006 ± 0.008	0.97 ± 0.03	1.02 ± 0.08%
Brain	VDM (Ho et al., 2022)	247.84	69.23	0.46 ± 0.03	0.89 ± 0.02	1.02 ± 0.07%
	CogVideoX (Yang et al., 2024)	302.20	114.50	0.46 ± 0.04	0.69 ± 0.27	1.06 ± 2.8%
	DynamiCrafter (Xing et al., 2025)	288.62	142.40	1.05 ± 0.03	0.88 ± 0.03	1.03 ± 0.20%
	IGG	26.39	6.23	0.006 ± 0.008	0.97 ± 0.03	1.02 ± 0.08%

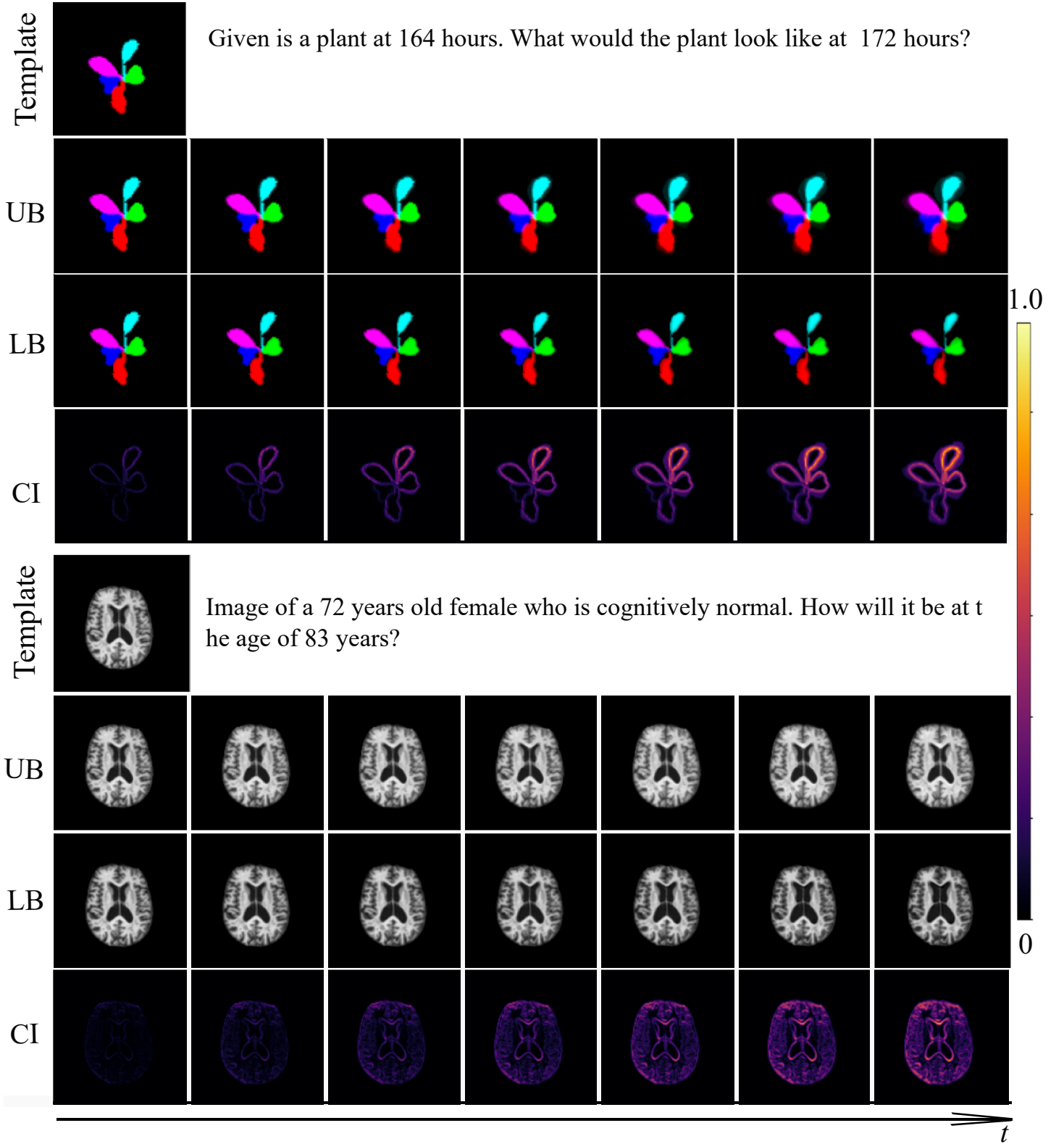


Figure 5: Top: Input template images with text instructions. Bottom: Confidence maps illustrating the lower bounds (LB), upper bounds (UB), and confidence intervals (CI), which represent regions representing 95% of ideal growth patterns, based on 1000 samples generated by our IGG model across different time frames.

5. Conclusion & Discussion

This paper presents an extended work, IGG (Wu et al., 2025), a geodesic-informed generative diffusion model for

topology-preserved image video generation. In contrast to existing generative models that focus primarily on manipulating image intensity and texture, IGG explicitly learns the underlying latent distribution of geodesic-informed deforma-

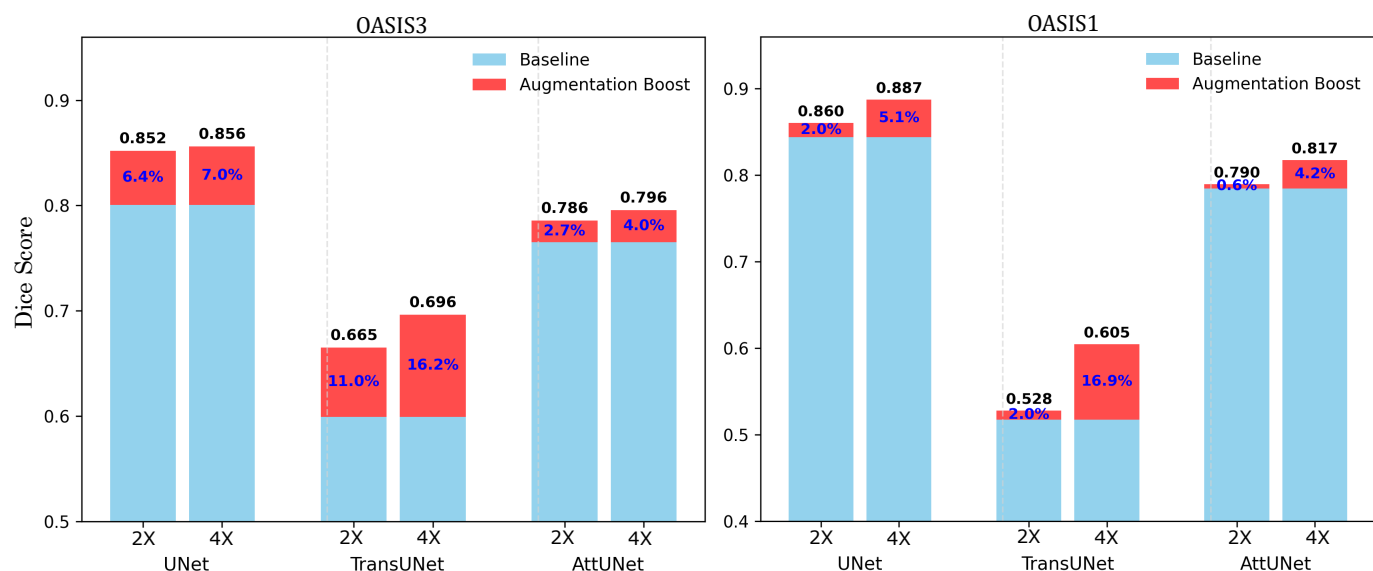


Figure 6: The improvement in average dice scores achieved by different segmentation models when trained with varying training dataset configurations. Baseline denotes training with the original limited dataset, while $2\times$ and $4\times$ correspond to augmentations of 2 times and 4 times the size of the original dataset, respectively.

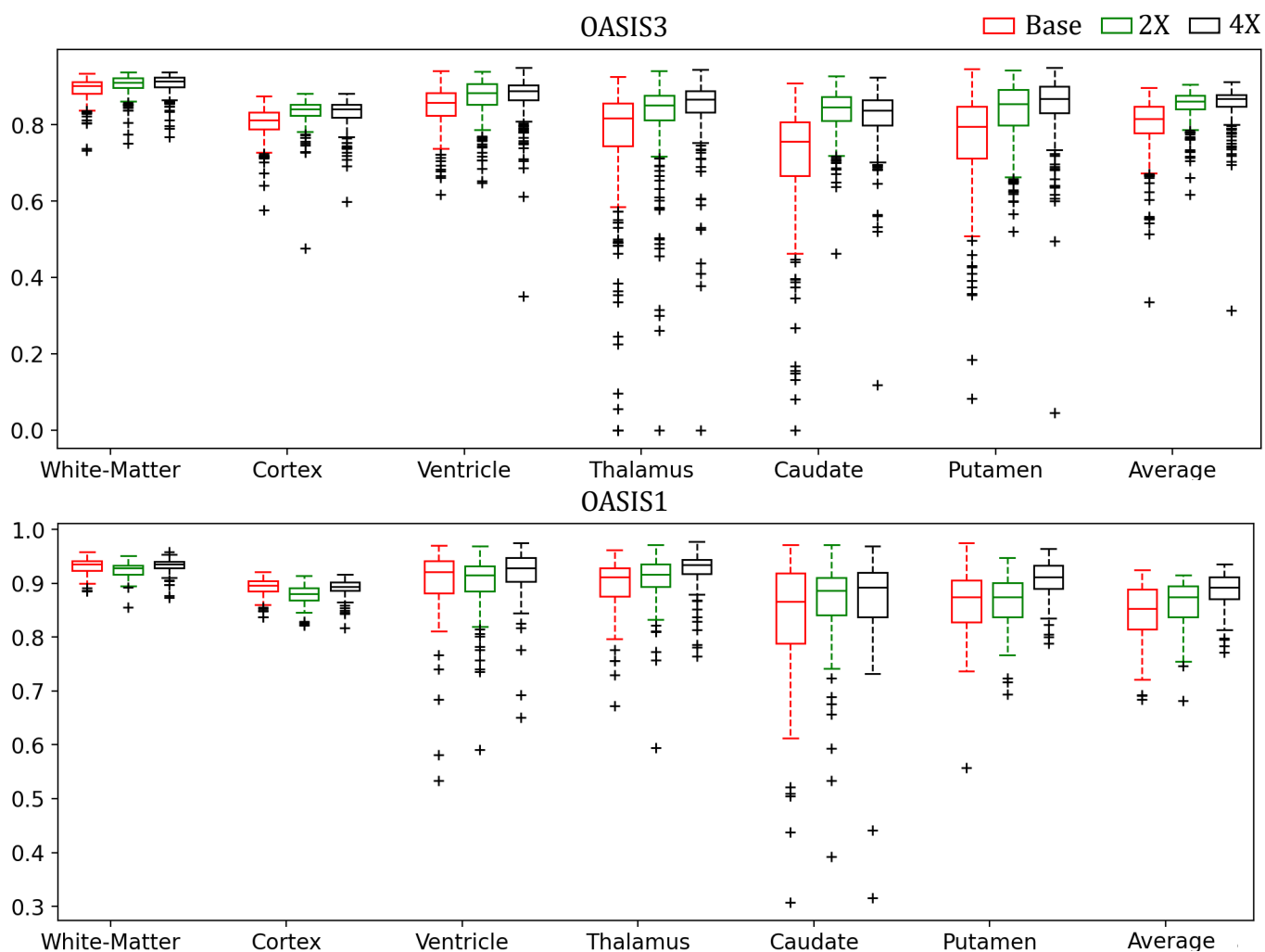


Figure 7: A comparison of average dice scores across different anatomical structures under varying training dataset configurations. Left to right: **base**: training with the original limited dataset, and $2\times$ and $4\times$ correspond to augmented sizes of the original dataset. Top to bottom: results on the OASIS-3 vs. OASIS-1.

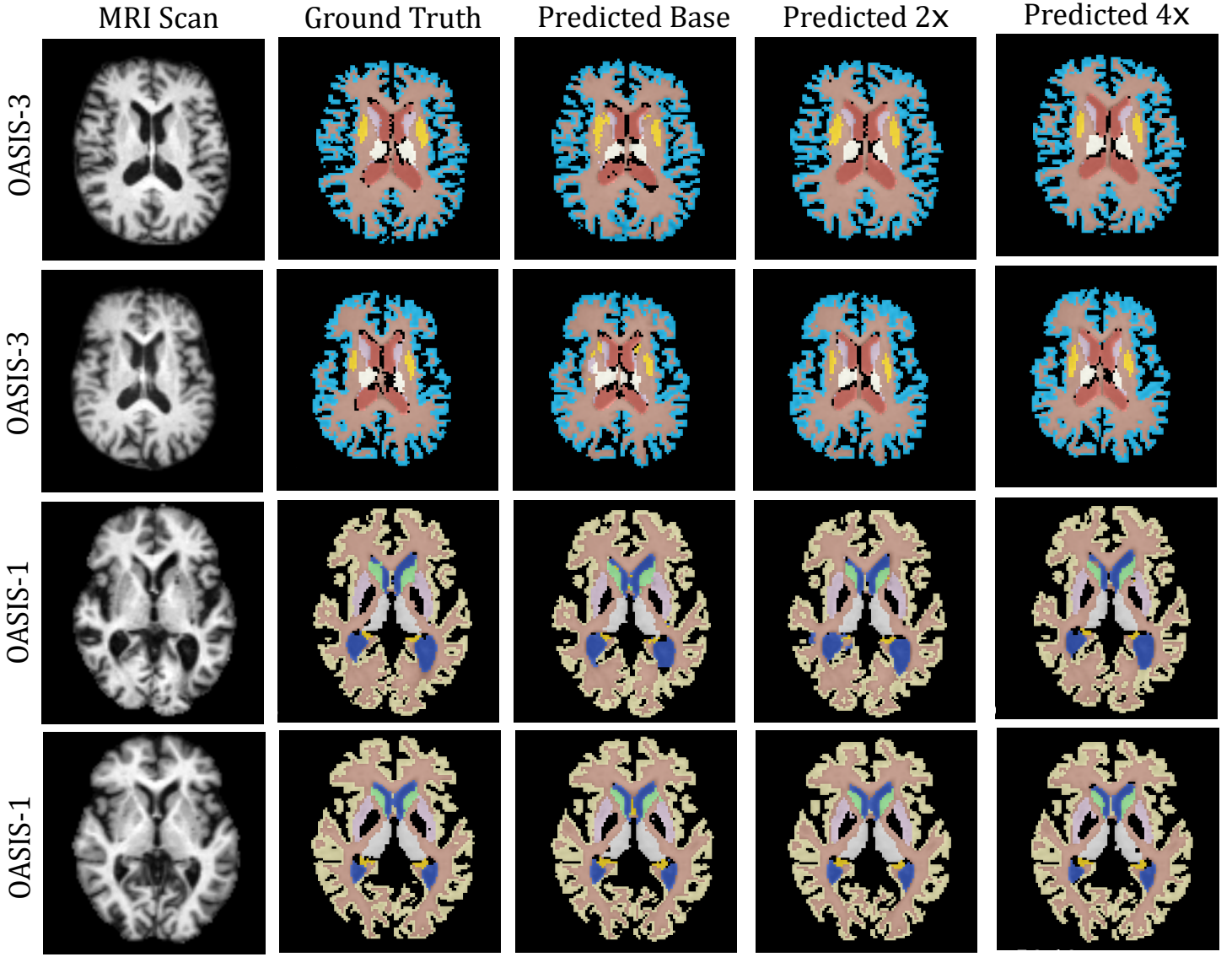


Figure 8: Examples of segmentation results on brain MRIs across different anatomical structures under varying training dataset configurations. Left to right: original MRI scans, ground truth labels, predicted labels from a model trained on the original limited dataset, and predicted labels from models trained with $2\times$ and $4\times$ the size of the original dataset. Top to bottom: results on the OASIS-3 vs. OASIS-1.

tions from images, enabling the synthesis of diverse samples within geodesic deformation spaces. Our work is the first generative model to operate on natural and interpretable image transformations: it not only captures shape-changing processes during generation, but also provides quantitative metrics to assess topological consistency in the synthesized samples. The key contributions of IGG include (i) a geodesic-informed registration network that learns latent representations of time-dependent diffeomorphic transformations, enabling accurate modeling of complex shape dynamics; and (ii) a latent geometric diffusion model that captures sequential deformation features conditioned on text inputs and template images, allowing controlled and semantically meaningful generation. Our experimental results demonstrate that IGG significantly improves sample

fidelity and diversity compared to state-of-the-art methods. Additionally, downstream segmentation experiments show substantial performance gains when using IGG-generated samples as data augmentation, particularly in low-data regimes.

Our proposed IGG model advances image generation by integrating geodesic principles, providing tools for topology assessment, and establishing a framework for synthesizing anatomically consistent image samples. We acknowledge that the fidelity of the synthesized data from our IGG model is inherently dependent on the accuracy of the underlying registration. When registration errors are large, the quality of the synthesized results may be correspondingly limited. This dependency reflects a critical assumption of our framework that anatomically meaningful correspon-

dences can be reliably established. Our topological prior is most effective in applications where anatomical structures are expected to preserve diffeomorphic consistency. However, in practical settings, imaging noise, pathology-induced appearance changes, or other non-diffeomorphic structural changes may violate this assumption. In such cases, strict topology preservation may not fully capture the underlying anatomical variability. Our future work will be investigated to combine IGG's strength in learning deformation fields with the modeling of image texture distributions, which will enable the development of more powerful and realistic generative models that capture both structural and appearance variations. Other future research will focus on several potential directions in extending IGG to handle 4D data (3D + time) to model dynamic longitudinal anatomical changes, such as cardiac motion or brain changes over time. These directions aim to further enhance the clinical and scientific utility of generative modeling in medical imaging.

Acknowledgments

This work was supported by NSF CAREER Grant 2239977.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we do not have conflicts of interest.

Data availability

The Brain MRIs used in training and testing in this paper are from the Open Access Series of Imaging Studies (OASIS-3) dataset (LaMontagne et al., 2019), which are readily accessible and user-friendly. Readers interested in evaluating the accuracy of the method can access and utilize the publicly available datasets. The code for this study is also publicly available at <https://github.com/nellie689/IGG>.

References

Vladimir Arnold. Sur la géométrie différentielle des groupes de lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. In *Annales de l'institut Fourier*, volume 16, pages 319–361, 1966.

Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv:2304.08466*, 2023.

Duhyeon Bang and Hyunjung Shim. Mggan: Solving mode collapse using manifold-guided training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2347–2356, 2021.

Mirza Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv:1801.01401*, 2018.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Duy-Phuong Dao, Hyung-Jeong Yang, and Jahae Kim. Conditional diffusion model for longitudinal medical image generation. *arXiv preprint arXiv:2411.05860*, 2024.

Swakshar Deb, Nian Wu, Frederick H Epstein, and Miaomiao Zhang. Unsupervised cardiac video translation via motion feature guided diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 648–658. Springer, 2025.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Robert Graf, Joachim Schmitt, Sarah Schlaeager, Hendrik Kristian Möller, Vasiliki Sideri-Lampretsa, Anjany

- Sekuboyina, Sandro Manuel Krieg, Benedikt Wiestler, Bjoern Menze, Daniel Rueckert, et al. Denoising diffusion-based mri to ct image translation enables automated spinal segmentation. *European Radiology Experimental*, 7(1):70, 2023.
- Saumya Gupta, Dimitris Samaras, and Chao Chen. Topodiffusionnet: A topology-aware diffusion model. *arXiv:2410.16646*, 2024.
- Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- J. Hinkle, D. Womble, and H.J.d Yoon. Diffeomorphic autoencoders for LDDMM atlas building. In *Medical Imaging with Deep Learning*, 2018.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.
- Tonmoy Hossain and Miaomiao Zhang. Mgaug: Multi-modal geometric augmentation in latent spaces of image deformations. *Medical Image Analysis*, 102:103540, 2025.
- Jiangbei Hu, Ben Fei, Baixin Xu, Fei Hou, Weidong Yang, Shengfa Wang, Na Lei, Chen Qian, and Ying He. Topology-aware latent diffusion for 3d shape generation. *arXiv:2401.17603*, 2024.
- Nivetha Jayakumar, Tonmoy Hossain, and Miaomiao Zhang. Sadir: shape-aware diffusion models for 3d image reconstruction. In *International workshop on shape in medical imaging*, pages 287–300. Springer, 2023.
- Nivetha Jayakumar, Srivardhan Reddy Gadila, Tonmoy Hossain, Yangfeng Ji, and Miaomiao Zhang. Tpie: Topology-preserved image editing with text instructions. *arXiv preprint arXiv:2411.16714*, 2024.
- Sarang Joshi, Brad Davis, Matthieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 2004.
- Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.
- Boah Kim, Inhwa Han, and Jong Chul Ye. Diffusemorph: Unsupervised deformable image registration using diffusion model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 347–364. Springer, 2022.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv*, pages 2019–12, 2019.
- Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024.
- Wei Li, Li Fan, Zhenyu Wang, Chao Ma, and Xiaohui Cui. Tackling mode collapse in multi-generator gans with orthogonal vectors. *Pattern Recognition*, 110:107646, 2021.
- Xiaofeng Liu, Fangxu Xing, Maureen Stone, Jiachen Zhuo, Timothy Reese, Jerry L Prince, Georges El Fakhri, and Jonghye Woo. Generative self-training for cross-domain unsupervised tagged-to-cine mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 138–148. Springer, 2021.
- Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

- François Mazé and Faez Ahmed. Diffusion models beat gans on topology optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 9108–9116, 2023.
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv:2108.01073*, 2021.
- Michael I Miller, Alain Trouvé, and Laurent Younes. On the metrics and euler–lagrange equations of computational anatomy. *Annual Review of Biomedical Engineering*, 4(1):375–405, 2002.
- Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification, 2023.
- Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems*, 36, 2024.
- Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Öztürk, Alper Güngör, and Tolga Cukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023.
- Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23051–23061, 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- Jacob C Reinhold, Aaron Carass, and Jerry L Prince. A structural causal model for mr images of multiple sclerosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 782–792. Springer, 2021.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- Gulshan Sharma, Abhinav Dhall, and Ramanathan Subramanian. Medic: Mitigating eeg data scarcity via class-conditioned diffusion model. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.
- Sophie Starck, Vasiliki Sideri-Lampretsa, Bernhard Kainz, Martin J Menten, Tamara T Mueller, and Daniel Rueckert. Diff-def: Diffusion-generated deformation fields for conditional atlases. *IEEE Transactions on Medical Imaging*, 2025.
- Kerem C Tezcan, Christian F Baumgartner, Roger Luechinger, Klaas P Pruessmann, and Ender Konukoglu. Mr image reconstruction using deep density priors. *IEEE transactions on medical imaging*, 38(7):1633–1642, 2018.
- Hideaki Uchiyama, Shunsuke Sakurai, Masashi Mishima, Daisaku Arita, Takashi Okayasu, Atsushi Shimada, and Rin-ichiro Taniguchi. An easy-to-setup 3d phenotyping platform for komatsuna dataset. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2038–2045, 2017.
- Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv:1812.01717*, 2018.
- François X Vialard, Laurent Risser, Daniel Rueckert, and Colin J Cotter. Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation. *International Journal of Computer Vision*, 97(2):229–241, 2012.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- Nian Wu and Miaomiao Zhang. Neurepdiff: Neural operators to predict geodesics in deformation spaces. In *International Conference on Information Processing in Medical Imaging*, pages 588–600. Springer, 2023.

- Nian Wu and Miaomiao Zhang. Learning geodesics of geometric shape deformations from images. *The Journal of Machine Learning for Biomedical Imaging*, 2024.
- Nian Wu, Jiarui Xing, and Miaomiao Zhang. Tlrn: Temporal latent residual networks for large deformation image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 728–738. Springer, 2024.
- Nian Wu, Nivetha Jayakumar, Jiarui Xing, and Miaomiao Zhang. Igg: Image generation informed by geodesic dynamics in deformation spaces. In *International Conference on Information Processing in Medical Imaging*, pages 232–246. Springer, 2025.
- Jianhao Xie, Ziang Zhang, Zhenyu Weng, Yuesheng Zhu, and Guibo Luo. Meddiff-ft: Data-efficient diffusion model fine-tuning with structural guidance for controllable medical image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 306–316. Springer, 2025.
- Jiarui Xing, Nivetha Jayakumar, Nian Wu, Yu Wang, Frederick H Epstein, and Miaomiao Zhang. Lamod: Latent motion diffusion model for myocardial strain generation. In *International Workshop on Shape in Medical Imaging*, pages 164–177. Springer, 2024.
- Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025.
- Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 255–263. Springer, 2019.
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv:2408.06072*, 2024.
- Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging*, pages 388–400. Springer, 2023.
- Laurent Younes, Felipe Arrate, and Michael I Miller. Evolutions equations in computational anatomy. *NeuroImage*, 45(1):S40–S50, 2009.
- Zhengming Yu, Zhiyang Dou, Xiaoxiao Long, Cheng Lin, Zekun Li, Yuan Liu, Norman Müller, Taku Komura, Marc Habermann, Christian Theobalt, et al. Surf-d: Generating high-quality surfaces of arbitrary topologies using diffusion models. In *European Conference on Computer Vision*, pages 419–438. Springer, 2025.
- Kai Zhang, Geng Chen, Shijie Huang, Fangmei Zhu, Zhongxiang Ding, and Dinggang Shen. Development-driven diffusion model for longitudinal prediction of fetal brain mri with unpaired data. *IEEE Transactions on Medical Imaging*, 2024.
- Miaomiao Zhang, Ruizhi Liao, Adrian V Dalca, Esra A Turk, Jie Luo, P Ellen Grant, and Polina Golland. Frequency diffeomorphisms for efficient image registration. In *International Conference on Information Processing in Medical Imaging*, pages 559–570. Springer, 2017.
- Zhaoyu Zhang, Mengyan Li, and Jun Yu. On the convergence and mode collapse of gan. In *SIGGRAPH Asia 2018 Technical Briefs*, pages 1–4. 2018.